Evaluating Text-to-Visual Generation with Image-to-Text Generation

Year: 2024 | Citations: 321 | Authors: Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia

Abstract

Despite significant progress in generative AI, comprehensive evaluation remains challenging because of the lack of effective metrics and standardized benchmarks. For instance, the widely-used CLIPScore measures the alignment between a (generated) image and text prompt, but it fails to produce reliable scores for complex prompts involving compositions of objects, attributes, and relations. One reason is that text encoders of CLIP can notoriously act as a"bag of words", conflating prompts such as"the horse is eating the grass"with"the grass is eating the horse". To address this, we introduce the VQAScore, which uses a visual-question-answering (VQA) model to produce an alignment score by computing the probability of a"Yes"answer to a simple"Does this figure show '{text}'?"question. Though simpler than prior art, VQAScore computed with off-the-shelf models produces state-of-the-art results across many (8) image-text alignment benchmarks. We also compute VQAScore with an in-house model that follows best practices in the literature. For example, we use a bidirectional image-question encoder that allows image embeddings to depend on the question being asked (and vice versa). Our in-house model, CLIP-FlanT5, outperforms even the strongest baselines that make use of the proprietary GPT-4V. Interestingly, although we train with only images, VQAScore can also align text with video and 3D models. VQAScore allows researchers to benchmark text-to-visual generation using complex texts that capture the compositional structure of real-world prompts. We introduce GenAI-Bench, a more challenging benchmark with 1,600 compositional text prompts that require parsing scenes, objects, attributes, relationships, and high-order reasoning like comparison and logic. GenAI-Bench also offers over 15,000 human ratings for leading image and video generation models such as Stable Diffusion, DALL-E 3, and Gen2.